# UNITED STATES PATENT APPLICATION

of

EUGENE A. FITZGERALD

and

NICOLE GERRISH

for

# METHOD OF FABRICATING CMOS INVERTER AND INTEGRATED CIRCUITS UTILIZING STRAINED SILICON SURFACE CHANNEL MOSFETS

# METHOD OF FABRICATING CMOS INVERTER AND INTEGRATED CIRCUITS UTILIZING STRAINED SILICON SURFACE CHANNEL MOSFETS

## PRIORITY INFORMATION

This application claims priority from provisional application Ser. No. 60/ 250,985 filed December 4, 2000.

## BACKGROUND OF THE INVENTION

The invention relates to the field of strained silicon surface channel MOSFETs, and in particular to using them in CMOS inverters and other integrated circuits.

The ability to scale CMOS devices to smaller and smaller dimensions has enabled integrated circuit technology to experience continuous performance enhancement. Since the 1970's, gate lengths have decreased by two orders of magnitude, resulting in a 30% improvement in the price/performance per year. Historically, these gains have been dictated by the advancement of optical photolithography tools and photoresist materials. As CMOS device size progresses deeper and deeper into the sub-micron regime, the associated cost of these new tools and materials can be prohibitive. A state of the art CMOS facility can cost more than 1-2 billion dollars, a daunting figure considering that the lithography equipment is generally only useful for two scaling generations.

In addition to economic constraints, scaling is quickly approaching constraints of device materials and design. Fundamental physical limits such as gate oxide leakage and source/drain extension resistance make continued minimization beyond 0.1μm difficult if not impossible to maintain. New materials such as high k dielectrics and metal gate electrodes must be

introduced in order to sustain the current roadmap until 2005. Beyond 2005, the fate of scaling is unclear.

Since the limits of scaling are well within sight, researchers have actively sought other methods of increasing device performance. One alternative is to make heterostructure FETs in GaAs/AlGaAs in order to take advantage of the high electron mobilities in these materials. However, the high electron mobility in GaAs is partially offset by the low hole mobility, causing a problem for complementary FET architectures. In addition, GaAs devices are usually fabricated with Schottky gates. Schottky diodes have leakage currents that are orders of magnitudes higher than MOS structures. The excess leakage causes an increase in the off-state power consumption that is unacceptable for highly functional circuits. Schottky diodes also lack the self-aligned gate technology enjoyed by MOS structures and thus typically have larger gate-to-source and gate-to-drain resistances. Finally, GaAs processing does not enjoy the same economies of scale that have caused silicon technologies to thrive. As a result, wide-scale production of GaAs circuits would be extremely costly to implement.

The most popular method to increase device speed at a constant gate length is to fabricate devices on silicon-on-insulator (SOI) substrates. In an SOI device, a buried oxide layer prevents the channel from fully depleting. Partially depleted devices offer improvements in the junction area capacitance, the device body effect, and the gate-to-body coupling. In the best-case scenario, these device improvements will result in an 18% enhancement in circuit speed. However, this improved performance comes at a cost. The partially depleted floating body causes an uncontrolled lowering of the threshold voltage, known as the floating body effect. This phenomenon increases the off-state leakage of the transistor and thus offsets some

of the potential performance advantages. Circuit designers must extract enhancements through design changes at the architectural level. This redesign can be costly and thus is not economically advantageous for all Si CMOS products. Furthermore, the reduced junction capacitance of SOI devices is less important for high functionality circuits where the interconnect capacitance is dominant. As a result, the enhancement offered by SOI devices is limited in its scope.

Researchers have also investigated the mobility enhancement in strained silicon as a method to improve CMOS performance. To date, efforts have focused on circuits that employ a buried channel device for the PMOS, and a surface channel device for the NMOS. This method provides the maximum mobility enhancement; however, at high fields the buried channel device performance is complex due to the activation of two carrier channels. In addition, monolithic buried and surface channel CMOS fabrication is more complex than bulk silicon processing. This complexity adds to processing costs and reduces the device yield.

## SUMMARY OF THE INVENTION

In accordance with the invention, the performance of a silicon CMOS inverter by increasing the electron and hole mobilities is enhanced. This enhancement is achieved through surface channel, strained-silicon epitaxy on an engineered SiGe/Si substrate. Both the n-type and p-type channels (NMOS and PMOS) are surface channel, enhancement mode devices. The technique allows inverter performance to be improved at a constant gate length without adding complexity to circuit fabrication or design.

When silicon is placed under tension, the degeneracy of the conduction band splits

forcing two valleys to be occupied instead of six. As a result, the in-plane, room temperature electron mobility is dramatically increased, reaching a value as high as 2900 $cm^2$/V-sec in buried channel devices for electrons densities of $10^{11}$-$10^{12}cm^{-2}$. Mobility enhancement can be incorporated into a MOS device through the structure of the invention. In the structure, a compositionally graded buffer layer is used to accommodate the lattice mismatch between a relaxed SiGe film and a Si substrate. By spreading the lattice mismatch over a distance, the graded buffer minimizes the number of dislocations reaching the surface and thus provides a method for growing high-quality relaxed SiGe films on Si. Subsequently, a silicon film below the critical thickness can be grown on the SiGe film. Since the lattice constant of SiGe is larger than that of Si, the Si film is under biaxial tension and thus the carriers exhibit strain-enhanced mobilities.

There are two primary methods of extracting performance enhancement from the increased carrier mobility. First, the frequency of operation can be increased while keeping the power constant. The propagation delay of an inverter is inversely proportional to the carrier mobility. Thus, if the carrier mobility is increased, the propagation delay decreases, causing the overall device speed to increase. This scenario is useful for applications such as desktop computers where the speed is more crucial than the power consumption. Second, the power consumption can be decreased at a constant frequency of operation. When the carrier mobility increases, the gate voltage can be reduced by an inverse fraction while maintaining the same inverter speed. Since power is proportional to the square of the gate voltage, this reduction results in a significant decrease in the power consumption. This situation is most useful for portable applications that operate off of a limited power supply.

Unlike GaAs high mobility technologies, strained silicon devices can be fabricated with standard silicon CMOS processing methods and tools. This compatibility allows for performance enhancement with no additional capital expenditures. The technology is also scalable and thus can be implemented in both long and short channel devices. The physical mechanism behind short channel mobility enhancement is not completely understood; however it has been witnessed and thus can be used to improve device performance. Furthermore, if desired, strained silicon can be incorporated with SOI technology in order to provide ultra-high speed/low power circuits. In summary, since strained silicon technology is similar to bulk silicon technology, it is not exclusive to other enhancement methods. As a result, strained silicon is an excellent technique for CMOS performance improvement.

## BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 is a cross-section of the substrate structure required to produce a strained silicon surface channel MOSFET;

Figs. 2A and 2B are graphs of mobility enhancements for electrons and holes, respectively, for strained silicon on $Si_{1-x}Ge_x$ for x=10-30%;

Fig. 3 is a table that displays surface roughness data for various relaxed SiGe buffers on Si substrates;

Fig. 4 is a schematic diagram of a CMOS inverter;

Figs. 5A and 5B are schematic diagrams of the structures of a strained silicon MOSFET 500 and a strained silicon MOSFET 550 on SOI, respectively;

Fig. 6 is a table showing electron and hole mobility enhancements measured for

strained silicon on 20% and 30% SiGe;

Fig. 7 is a table showing inverter characteristics for 1.2μm CMOS fabricated in both bulk and strained silicon when the interconnect capacitance is dominant;

Fig. 8 is a table showing additional scenarios for strained silicon inverters when the interconnect capacitance is dominant;

Fig. 9 is a table showing inverter characteristics for 1.2μm CMOS fabricated in both bulk and strained silicon when the device capacitance is dominant;

Fig. 10 is a graph showing NMOSFET transconductance versus channel length for various carrier mobilities;

Fig. 11 is a graph showing the propagation delay of a 0.25μm CMOS inverter for a range of electron and hole mobility enhancements;

Figs. 12A-12E show a fabrication process sequence for strained silicon on SOI substrates; and

Figs. 13A-13C are circuit schematics for a NOR gate, a NAND gate and a XOR gate, respectively.

## DETAILED DESCRIPTION OF THE INVENTION

*Strained Silicon Enhancement*

Fig. 1 is a cross-section of the substrate structure 100 required to produce a strained silicon surface channel MOSFET. The larger lattice constant, relaxed SiGe layer applies biaxial strain to the silicon surface layer. In this structure, a compositionally graded buffer layer 102 is used to accommodate the lattice mismatch between a relaxed SiGe film 106 and a

Si substrate 104. By spreading the lattice mismatch over a distance, the graded buffer minimizes the number of dislocations reaching the surface and thus provides a method for growing high-quality relaxed SiGe films on Si. Subsequently, a silicon film 108 below the critical thickness can be grown on the SiGe film. Since the lattice constant of SiGe is larger than that of Si, the Si film is under biaxial tension and thus the carriers exhibit strain-enhanced mobilities. Thereafter, a layer 110 of $SiO_2$ and a gate 112 are provided thereon.

In the structure shown in Fig.1, the silicon channel is placed under biaxial tension by the underlying, larger lattice constant SiGe layer. This strain causes the conduction band to split into two-fold and four-fold degenerate bands. The two-fold band is preferentially occupied since it sits at a lower energy. The energy separation between the bands is approximately

$$\Delta E_{strain} = 0.67 \cdot x \text{ (eV)} \quad (1)$$

where x is equal to the Ge content in the SiGe layer. The equation shows that the band splitting increases as the Ge content increases. This splitting causes mobility enhancement by two mechanisms. First, the two-fold band has a lower effective mass, and thus higher mobility than the four-fold band. Therefore, as the higher mobility band becomes energetically preferred, the average carrier mobility increases. Second, since the carriers are occupying two orbitals instead of six, inter-valley phonon scattering is reduced, further enhancing the carrier mobility.

The effects of Ge concentration on electron and hole mobility for a surface channel device can be seen in Figs. 2A and 2B, respectively. Figs. 2A and 2B are graphs of mobility enhancements for electrons and holes, respectively, for strained silicon on $Si_{1-x}Ge_x$ for x=10-

30%. At 20% Ge, the electron enhancement at high fields is approximately 1.75 while the hole enhancement is essentially negligible. Above approximately 20% Ge, the electron enhancement saturates. This saturation occurs because the conduction band splitting is large enough that almost all of the electrons occupy the high mobility band. Hole enhancement saturation has not yet been observed; therefore, raising the Ge concentration to 30% increases hole mobility by a factor of 1.4. Hole enhancement saturation is predicted to occur at a Ge concentration of about 40%.

The low hole mobility in surface channel devices has caused other researchers to move to higher mobility, buried channel devices for the PMOSFET. Here, it is shown that significant CMOS enhancement can be achieved using surface channel devices for both NMOS and PMOS. This design allows for high performance without the complications of dual channel operation and without adding complexity to circuit fabrication.

Until recently, the material quality of relaxed SiGe on Si was insufficient for utilization in CMOS fabrication. During epitaxial growth, the surface of the SiGe becomes very rough as the material is relaxed via dislocation introduction. Researchers have tried to intrinsically control the surface morphology through the growth; however, since the stress fields from the misfit dislocations affect the growth front, no intrinsic epitaxial solution is possible. U.S. Pat. No. 6,107,653 issued to Fitzgerald, incorporated herein by reference, describes a method of planarization and regrowth that allows all devices on relaxed SiGe to possess a significantly flatter surface. This reduction in surface roughness is critical in the production of strained Si CMOS devices since it increases the yield for fine-line lithography.

Fig. 3 is a table that displays surface roughness data for various relaxed SiGe buffers

on Si substrates. It will be appreciated that the as-grown crosshatch pattern for relaxed $Si_{0.8}Ge_{0.2}$ buffers creates a typical roughness of approximately 7.9nm. This average roughness increases as the Ge content in the relaxed buffer is increased. Thus, for any relaxed SiGe layer that is relaxed through dislocation introduction during growth, the surface roughness is unacceptable for state-of-the-art fabrication facilities. After the relaxed SiGe is planarized, the average roughness is less than 1nm (typically 0.57nm), and after a 1.5μm device layer deposition, the average roughness is 0.77nm. Therefore, after the complete structure is fabricated, there is over an order of magnitude reduction in the surface roughness. The resulting high quality material is well suited for state of the art CMOS processing.

*CMOS Inverter*

Fig. 4 is a schematic diagram of a CMOS inverter 400. When the input voltage, $V_{in}$, to the inverter is low, a PMOS transistor 402 turns on, charges up a load capacitance 404, and the output goes to a gate drive 406, $V_{DD}$. Alternatively, when $V_{in}$ is high, an NMOS transistor 408 turns on, discharges the load capacitance, and the output node goes to ground 410. In this manner, the inverter is able to perform the logic swing necessary for digital processing. The load capacitance, denoted as $C_L$, represents a lumped model of all of the capacitances between $V_{out}$ and ground.

Since the load capacitance must be fully charged or discharged before the logic swing is complete, the magnitude of $C_L$ has a large impact on inverter performance. The performance is usually quantified by two variables: the propagation delay, $t_p$, and the power consumed, P. The propagation delay is defined as how quickly a gate responds to a change in its input and is

given by

$$t_p = \frac{C_L \cdot V_{DD}}{I_{av}} \quad (2)$$

where $I_{av}$ is the average current during the voltage transition. There is a propagation delay term associated with the NMOS discharging current, $t_{pHL}$, and a term associated with the PMOS charging current, $t_{pLH}$. The average of these two values represents the overall inverter delay:

$$t_p = \frac{t_{pHL} + t_{pLH}}{2} \quad (3)$$

Assuming that static and short-circuit power are negligible, the power consumed can be written as

$$P = \frac{C_L \cdot V_{DD}^{\ 2}}{t_p} \quad (4)$$

From equations 2 and 4, one can see that both the propagation delay and the power consumption have a linear dependence on the load capacitance. In an inverter, $C_L$ consists of two major components: interconnect capacitance and device capacitance. Which component dominates $C_L$ depends on the architecture of the circuit in question.

*Strained Silicon, Long Channel CMOS Inverter*

Figs. 5A and 5B are schematic diagrams of the structures of a strained silicon MOSFET 500 and a strained silicon MOSFET 550 on SOI, respectively. The structure in Fig. 5A contains the elements shown in the substrate structure of Fig. 1 along with basic elements

of the MOSFET device structure, i.e. source 513 and drain 514 regions, gate oxide 510 and gate 512 layers, and device isolation regions 516. Figure 5B shows the same device elements on a SiGe-on-insulator (SGOI) substrate. In the SGOI substrate, a buried oxide layer 518 separates the relaxed SiGe layer 506 from the underlying Si substrate 504. In both MOSFET structures, the strained Si layer 508 serves as the carrier channel, thus enabling improved device performance over their bulk Si counterparts.

When strained silicon is used as the carrier channel, the electron and hole mobilities are multiplied by enhancement factors. Figs. 2A and 2B demonstrate that this enhancement differs for electrons and holes and also that it varies with the Ge fraction in the underlying SiGe layer.

A summary of the enhancements for $Si_{0.8}Ge_{0.2}$ and $Si_{0.7}Ge_{0.3}$ is shown in Fig. 6. Fig. 6 is a table showing electron and hole mobility enhancements measured for strained silicon on 20% and 30% SiGe. These enhancements are incorporated into 1.2μm CMOS models in order to quantify the effects on inverter performance. The mobility enhancement can be capitalized upon in two primary ways: 1) increase the inverter speed at a constant power and 2) reduce the inverter power at a constant speed. These two optimization methods are investigated for both a wiring capacitance dominated case and a device capacitance dominated case.

*Interconnect Dominated Capacitance*

In high performance microprocessors, the interconnect or wiring capacitance is often dominant over the device capacitance. In this scenario, standard silicon PMOS devices are made two to three times wider than their NMOS counterparts. This factor comes from the

ratio of the electron and hole mobilities in bulk silicon. If the devices were of equal width, the low hole mobility would cause the PMOS device to have an average current two to three times lower than the NMOS device. Equation 2 shows that this low current would result in a high $t_{pLH}$ and thus cause a large gate delay. Increasing the width of the PMOS device equates the high-to-low and low-to-high propagation delays and thus creates a symmetrical, high-speed inverter.

Key values for a bulk silicon, 1.2μm symmetrical inverter are shown in Fig. 7. Fig. 7 is a table showing inverter characteristics for 1.2μm CMOS fabricated in both bulk and strained silicon when the interconnect capacitance is dominant. The strained silicon inverters are optimized to provide high speed at constant power and low power at constant speed. The propagation delay for the bulk silicon inverter is 204 psec and the consumed power is 3.93mW. In an application where speed is paramount, such as in desktop computing, strained silicon provides a good way to enhance the circuit speed. Assuming no change from the bulk silicon design, a strained silicon inverter on $Si_{0.8}Ge_{0.2}$ results in a 15% speed increase at constant power. When the channel is on $Si_{0.7}Ge_{0.3}$, the speed enhancement improves to 29% (Fig. 7).

The improvement in inverter speed expected with one generation of scaling is approximately 15% (assumes an 11% reduction in feature size). Thus, the speed enhancement provided by a strained silicon inverter on 20% SiGe is equal to one scaling generation, while the speed enhancement provided by 30% SiGe is equivalent to two scaling generations.

Alternatively, reducing the gate drive, $V_{DD}$, can reduce the power at a constant speed. For 20% SiGe, the power consumption is 27% lower than its bulk silicon counterpart. When

30% SiGe is used, the power is reduced by 44% from the bulk silicon value (Fig. 7). This power reduction is important for portable computing applications such as laptops and handhelds.

Equation 4 shows that if $C_L$ is constant and $t_p$ is reduced, $V_{DD}$ must decrease to maintain the same inverter power. If the power consumption is not critical, the inverter frequency can be maximized by employing strained silicon devices at the same $V_{DD}$ as bulk Si devices. As described heretofore above, in a constant power scenario, the inverter speed is increased 15% for Si on $Si_{0.8}Ge_{0.2}$ and 29% for Si on $Si_{0.7}Ge_{0.3}$. When $V_{DD}$ is held constant, this enhancement increases to 29% and 58%, for Si on $Si_{0.8}Ge_{0.2}$ and $Si_{0.7}Ge_{0.3}$, respectively. Fig. 8 is a table showing additional scenarios for strained silicon inverters on 20% and 30% SiGe when the interconnect capacitance is dominant. Parameters are given for 1) strained silicon inverters with the same $V_{DD}$ as comparable bulk silicon inverters 2) symmetrical strained silicon inverters -designed for high speed and 3) symmetrical strained silicon inverters designed for low power.

One drawback of strained silicon, surface channel CMOS is that the electron and hole mobilities are unbalanced further by the uneven electron and hole enhancements. This unbalance in mobility translates to an unbalance in the noise margins of the inverter. The noise margins represent the allowable variability in the high and low inputs to the inverter. In bulk silicon microprocessors, both the low and high noise margins are about 2.06 V. For strained silicon on 20% and 30% SiGe, the low noise margin, $NM_L$, is decreased to 1.65 V and 1.72 V, respectively. While the $NM_L$ is reduced, the associated $NM_H$ is increased. Therefore, if the high input is noisier than the low input, the asymmetric noise margins may be

acceptable or even desired.

However, if a symmetrical inverter is required, the PMOS device width must be increased to $\mu_n/\mu_p$ times the NMOS device width. This translates to a 75% increase in PMOS width for $Si_{0.8}Ge_{0.2}$, and a 29% increase for $Si_{0.7}Ge_{0.3}$. If the circuit capacitance is dominated by interconnects, the increased device area will not cause a significant increase in $C_L$. As a result, if the increased area is acceptable for the intended application, inverter performance can be further enhanced. In the constant power scenario, the speed can now be increased by 37% for $Si_{0.8}Ge_{0.2}$ and by 39% for $Si_{0.7}Ge_{0.3}$. When the power is reduced for a constant frequency, a 50% and 52% reduction in consumed power is possible with 20% and 30% SiGe, respectively (Fig. 8). However, in many applications an increase in device area is not tolerable. In these situations if inverter symmetry is required, it is best to use strained silicon on 30% SiGe. Since the electron and hole enhancement is comparable on $Si_{0.7}Ge_{0.3}$, it is easier to trade-off size for symmetry to meet the needs of the application.

*Non-Interconnect Dominant Capacitance*

The device capacitance is dominant over the wiring capacitance in many analog applications. The device capacitance includes the diffusion and gate capacitance of the inverter itself as well as all inverters connected to the gate output, known as the fan-out. Since the capacitance of a device depends on its area, PMOS upsizing results in an increase in $C_L$. If inverter symmetry is not a prime concern, reducing the PMOS device size can increase the inverter speed. This PMOS downsizing has a negative effect on $t_{pLH}$ but has a positive effect on $t_{pHL}$. The optimum speed is achieved when the ratio between PMOS and NMOS widths is

set to $\sqrt{\mu_n / \mu_p}$ , where $\mu_n$ and $\mu_p$ represent the electron and hole mobilities, respectively. The optimized design has a propagation delay as much as 5% lower than the symmetrical design. The down side is that making $t_{pLH}$ and $t_{pHL}$ unbalanced reduces the low noise margin by approximately 15%. In most designs, this reduced $NM_L$ is still acceptable.

Fig. 9 is a table showing inverter characteristics for 1.2μm CMOS fabricated in both bulk and strained silicon when the device capacitance is dominant. The strained silicon inverters are optimized to provide high speed at constant power and low power at constant speed. For strained silicon on $Si_{0.8}Ge_{0.2}$, the electron mobility is a factor of 5.25 higher than the hole mobility. When the PMOS width is re-optimized to accommodate these mobilities, i.e., by using the $\sqrt{\mu_n / \mu_p}$ optimization, the strained silicon PMOS device on $Si_{0.8}Ge_{0.2}$ is over 30% wider than the bulk Si PMOS device.

The resulting increase in capacitance offsets some of the advantages of the enhanced mobility. Therefore, only a 4% speed increase occurs at constant power, and only an 8% decrease in power occurs at constant speed (Fig. 9). Although these improvements are significant, they represent a fraction of the performance improvement seen with a generation of scaling and do not surpass the performance capabilities available with SOI architectures.

In contrast, strained silicon on $Si_{0.7}Ge_{0.3}$ offers a significant performance enhancement at constant gate length for circuits designed to the $\sqrt{\mu_n / \mu_p}$ optimization. Since the electron and hole mobilities are more balanced, the effect on the load capacitance is less substantial. As a result, large performance gains can be achieved. At constant power, the inverter speed can be increased by over 23% and at constant speed, the power can be reduced by over 37% (Fig.

9). The latter enhancement has large implications for portable analog applications such as wireless communications.

As in the microprocessor case (interconnect dominated), the strained silicon devices suffer from small low noise margins. Once again, this effect can be minimized by using 30% SiGe. If larger margins are required, the PMOS device width can be increased to provide the required symmetry. However, this PMOS upsizing increases $C_L$ and thus causes an associated reduction in performance. Inverter design must be tuned to meet the specific needs of the intended application.

*Short Channel CMOS Inverter*

In short channel devices, the lateral electric field driving the current from the source to the drain becomes very high. As a result, the electron velocity approaches a limiting value called the saturation velocity, $v_{sat}$. Since strained silicon provides only a small enhancement in $v_{sat}$ over bulk silicon, researchers believed that strained silicon would not provide a performance enhancement in short channel devices. However, recent data shows that transconductance values in short channel devices exceed the maximum value predicted by velocity saturation theories. Fig. 10 is a graph showing NMOSFET transconductance versus channel length for various carrier mobilities. The dashed line indicates the maximum transconductance predicted by velocity saturation theories. The graph shows that high low-field mobilities translate to high high-field mobilities. The physical mechanism for this phenomenon is still not completely understood; however, it demonstrates that short channel mobility enhancement can occur in strained silicon.

The power consumed in an inverter depends on both $V_{DD}$ and $t_p$ (equation 4). Therefore, as $t_p$ is decreased due to mobility enhancement, $V_{DD}$ must also be decreased in order to maintain the same power consumption. In a long channel device, the average current, $I_{av}$, is proportional to $V_{DD}^2$. Inserting this dependence into equation 2 reveals an inverse dependence of the propagation delay on $V_{DD}$. Thus, as the average current in strained silicon is increased due to mobility enhancement, the effect on the propagation delay is somewhat offset by the reduction in $V_{DD}$.

A comparison of the high-speed scenario in Fig. 7 to the constant $V_{DD}$ scenario in Fig. 8 reveals the effect the reduced $V_{DD}$ has on speed enhancement. In a short channel device, the average current is proportional to $V_{DD}$ not $V_{DD}^2$, causing the propagation delay to have no dependence on $V_{DD}$ (assuming $V_{DD} >> V_T$). As a result, mobility enhancements in a short channel, strained silicon inverter are directly transferred to a reduction in $t_p$. A 1.2μm strained silicon inverter on 30% SiGe experiences a 29% increase in device speed for the same power. Assuming the same levels of enhancement, a short channel device experiences a 58% increase in device speed for constant power, double the enhancement seen in the long channel device.

Fig. 11 is a graph showing the propagation delay of a 0.25μm CMOS inverter for a range of electron and hole mobility enhancements. Although the exact enhancements in a short channel device vary with the fabrication processes, Fig. 11 demonstrates that even small enhancements can result in a significant effect on $t_p$.

*Strained Silicon on SOI*

Strained silicon technology can also be incorporated with SOI technology for added performance benefits. Figs. 12A-12E show a fabrication process sequence for strained silicon

on SOI substrates. First, a SiGe graded buffer layer 1202 is grown on a silicon substrate 1200 with a uniform relaxed SiGe cap layer 1204 of the desired concentration (Fig. 12A). This wafer is then bonded to a silicon wafer 1206 oxidized with a $SiO_2$ layer 1208 (Figs. 12B-12C). The initial substrate and graded layer are then removed through either wafer thinning or delamination methods. The resulting structure is a fully relaxed SiGe layer on oxide (Fig. 12D). A strained silicon layer 1210 can subsequently be grown on the engineered substrate to provide a platform for strained silicon, SOI devices (Fig. 12E). The resulting circuits would experience the performance enhancement of strained silicon as well as about an 18% performance improvement from the SOI architecture. In short channel devices, this improvement is equivalent to 3-4 scaling generations at a constant gate length.

A similar fabrication method can be used to provide relaxed SiGe layers directly on Si, i.e., without the presence of the graded buffer or an intermediate oxide. This heterostructure is fabricated using the sequence shown in Figs. 12A-12D without the oxide layer on the Si substrate. The graded composition layer possesses many dislocations and is quite thick relative to other epitaxial layers and to typical step-heights in CMOS. In addition, SiGe does not transfer heat as rapidly as Si. Therefore, a relaxed SiGe layer directly on Si is well suited for high power applications since the heat can be conducted away from the SiGe layer more efficiently.

*Other Digital Gates*

Although the preceding embodiments describe the performance of a CMOS inverter, strained silicon enhancement can be extended to other digital gates such as NOR, NAND, and

XOR structures. Circuit schematics for a NOR gate 1300, a NAND gate 1302 and a XOR gate 1304 are shown in Figs. 13A-C, respectively. The optimization procedures are similar to that used for the inverter in that the power consumption and/or propagation delay must be minimized while satisfying the noise margin and area requirements of the application. When analyzing these more complex circuits, the operation speed is determined by the worst-case delay for all of the possible inputs.

For example, in the pull down network of the NOR gate 1300 shown in Fig. 13A, the worst delay occurs when only one NMOS transistor is activated. Since the resistances are wired in parallel, turning on the second transistor only serves to reduce the delay of the network. Once the worst-case delay is determined for both the high to low and low to high transitions, techniques similar to those applied to the inverter can be used to determine the optimum design.

The enhancement provided by strained silicon is particularly beneficial for NAND-only architectures. As shown in Fig. 13B, in the architecture of the NAND gate 1302, the NMOS devices are wired in series while the PMOS devices are wired in parallel. This configuration results in a high output when either input A or input B is low, and a low output when both input A and input B are high, thus providing a NAND logic function. Since the NMOS devices are in series in the pull down network, the NMOS resistance is equal to two times the device resistance. As a result, the NMOS gate width must be doubled to make the high to low transition equal to the low to high transition.

Since electrons experience a larger enhancement than holes in strained Si, the NMOS gate width up scaling required in NAND-only architectures is less severe. For 1.2μm strained

silicon CMOS on a $Si_{0.8}Ge_{0.2}$ platform, the NMOS gate width must only be increased by 14% to balance the pull down and pull up networks (assuming the enhancements shown in Fig. 6). Correspondingly, for 1.2μm CMOS on $Si_{0.7}Ge_{0.3}$, the NMOS width must be increased by 55% since the n and p enhancements are more balanced. The high electron mobility becomes even more important when there are more than two inputs to the NAND gate, since additional series-wired NMOS devices are required.

Although the present invention has been shown and described with respect to several preferred embodiments thereof, various changes, omissions and additions to the form and detail thereof, may be made therein, without departing from the spirit and scope of the invention.

What is claimed is: